

The evolution of two-component systems in bacteria reveals different strategies for niche adaptation

Eric Alm^{1,2}, Katherine Huang^{1,3}, Adam Arkin^{1,3,4*}

1 The Virtual Institute for Microbial Stress and Survival, Berkeley, California, United States of America,

2 Massachusetts Institute of Technology, Department of Civil and Environmental Engineering and Biological Engineering, Cambridge, Massachusetts, United States of America,

3 Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, United States of America,

4 Department of Bioengineering, University of California Berkeley, Berkeley, California United States of America

Evolution of Histidine Protein Kinases

Editor: Philip E. Bourne, University of California San Diego, United States of America

Received December 30, 2005; **Accepted** September 19, 2006; **Published** , 2006

DOI: 10.1371/journal.pcbi.0020143

Copyright: © 2006 Alm et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations:

* To whom correspondence should be addressed. E-mail: aparkin@lbl.gov

 These authors contributed equally to this work.

Two-component systems including histidine protein kinases represent the primary signal transduction paradigm in prokaryotic organisms. To understand how these systems adapt to allow organisms to detect niche-specific signals, we analyzed the phylogenetic distribution of nearly 5000 histidine protein kinases from 207 sequenced prokaryotic genomes. We found that many genomes carry a large repertoire of recently evolved signaling genes, which may reflect selective pressure to adapt to new environmental conditions. Both lineage-specific gene family expansion and horizontal gene transfer play major roles in the introduction of new histidine kinases into genomes; however, there are differences in how these two evolutionary forces act. Genes imported *via* horizontal transfer are more likely to retain their original functionality as inferred from a similar complement of signaling domains, while gene family expansion accompanied by domain shuffling appears to be a major source of novel genetic diversity. Family expansion is the dominant source of new histidine kinase genes in the genomes most enriched in signaling proteins, and detailed analysis reveals that divergence in domain structure and changes in expression patterns are hallmarks of recent expansions. Finally, while these two modes of gene acquisition are widespread across bacterial taxa, there are clear species-specific preferences for which mode is used.

Citation: Alm E, Huang K, Arkin A (2006) The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. PLoS Comput Biol 2(10): e143. DOI: 10.1371/journal.pcbi.0020143

Synopsis

Pathways containing Histidine Protein Kinases (HPKs) represent a key mechanism for signal transduction, especially in bacteria. These systems help cells to sense and respond

to their environment by detecting external cues and effecting internal responses such as changes in gene expression. As such, they are believed to play a key role in niche adaptation, yet their evolution is difficult to study due to the large number of paralogous subfamilies. This work extends previous large-scale gene evolution studies by allowing complex paralogy relationships, and uncovers an abundance of horizontal transfers, gene duplications, and domain-shuffling that has marked the evolutionary history of HPKs. An important finding of this study is qualitative differences between the main strategies for acquiring new HPKs (HGT and gene duplication). Hallmarks of the latter process include domain-shuffling and the generation of ‘orphan’ HPKs not co-transcribed with a cognate response regulator.

Introduction

Bacteria change their physiological behavior according to signals detected in their environment. Typically, these changes are reflected in the alteration of gene expression patterns. These changes can be the result of action by a number of different types of signaling proteins including histidine protein kinases (HPKs) and their cognate response regulators (RRs), methyl-accepting chemotaxis proteins, di-guanylate and adenylate cyclases, Serine/Threonine/Tyrosine-protein kinases, as well as individual transcription factors or ‘one-component’ signal transduction proteins [1, 2]. Among these various protein families, HPKs are one of the most abundant, and historically have been regarded as the primary mechanism for signal transduction in bacteria [3].

Histidine kinases, and more generally signal transduction proteins, are thought to play a major role in the adaptation of bacteria to new or changing environments.

Consistent with this hypothesis, those bacteria that have the largest complements of signaling proteins generally tend to be bacteria with complex lifestyles such as *Myxococcus xanthus*, those that are found ubiquitously in varied environments such as *Pseudomonas*, or bacteria with numerous alternative metabolic strategies such as various δ - and ϵ -Proteobacteria [2, 4]. By contrast, few HPKs have been identified in the reduced genomes of parasitic bacteria, which likely have a relatively constant external environment.

While these signal transduction systems are thought to be a key part of the adaptive evolution of bacteria, few details are known about this process. In this study, we investigated the distribution of HPKs in sequenced bacterial genomes to address some fundamental questions: (i) What fraction of HPKs in a given genome represents newly acquired/ancient genes? (ii) What are the evolutionary processes that give rise to new HPKs? (iii) Do newly acquired HPKs sense similar signals or do they evolve new functionality?

We looked specifically at genes that entered into each lineage recently, making the logical assumption that recent additions are more likely to provide insight into the evolutionary basis of niche adaptation. Identifying recent acquisitions in a background of multiple paralogs is a difficult task. We describe a BLAST-based procedure for classifying and establishing the age of HPK domains. This procedure is derived from previous gene presence/absence studies by our group and others [5-7]. We based our phylogenetic analysis on the histidine kinase domain of each HPK only, allowing us to follow changes in the structure of the signaling domains (domain-shuffling) that generally lie upstream (N-terminal) to the kinase domain. The phylogenetic inference

procedure described requires an accurate species phylogeny, which we inferred from a concatenated gene profile including 15 ubiquitous bacterial genes without obvious paralogs (Supplementary Table S1).

We used the gene histories inferred by this procedure to estimate the relative contribution of horizontal gene transfer and gene duplication events to the evolution of new HPKs in each genome. We observed that some genomes acquired new HPKs primarily *via* horizontal gene transfer (HGT), while others relied mainly on lineage-specific expansion (LSE) of existing gene families. A closer look at genes acquired *via* these two mechanisms revealed differences in the extent to which upstream signaling domains and cognate response regulators were conserved as a result of each process, with HGT being more likely to preserve pre-existing relationships than gene duplication. We investigated one such HPK expansion in *Desulfovibrio vulgaris* in greater detail, and describe specific examples of LSE-associated domain shuffling. We further looked to functional genomic data and confirmed that these new HPKs have distinct gene expression profiles suggesting novel functional roles.

Inferring gene histories

In this study, we identified nearly 5000 HPKs from 207 genomes. We parsed each of these HPKs into domains (signaling domains of various types, histidine kinase domains, and response regulator domains in some cases), and analyzed the evolution of the histidine kinase domain from each gene using an approach based on pairwise BLASTp scores. In this way, we identified domains that were more similar to genes in the same genome than to genes in other genomes, as the likely result of lineage specific expansion (LSE). For domains that were more closely related to genes in distant

genomes than those in more closely related genomes, we inferred horizontal gene transfer (HGT) as a likely explanation. Finally, we identified a small number of subfamily ‘Birth’ events when a particular group of genes was found in a narrow range of species. These genes are sometimes referred to as ‘ORFans’ [6].

An overview of our approach is given in Figure 1. A complete outline of our algorithm for inferring domain histories is given in the Methods section, but a few key details are worth noting here. First, our approach relies on a species tree, which we constructed using a set of ubiquitous single-copy genes. This species tree is then condensed into a set of ‘outgroups’ at increasing evolutionary distances from the species of interest [5, 6, 8] as in Figure 2e. If the best hit of a HPK domain to a distant outgroup is closer than its best hit to a more closely-related outgroup, then it is considered to be ‘absent’ from the closer outgroup. When a gene is absent from two or more consecutive outgroups, implying multiple deletion events that could be alternatively explained by a single HGT event, the HGT event is inferred as more parsimonious, but *only if* the last outgroup before the presumed HGT diverged more recently than the phylogenetic cutoff distance we chose to define ‘recent’ events (see Methods). As noted later in the text, we reproduced the main findings described in the Results with a more conservative definition of HGT (absence from 3 consecutive outgroups) with nearly identical results. When the distance to a HPK domain within the same genome was closer than to the best hit from an outgroup, then a duplication (LSE) event was inferred, but *only if* the outgroup diverged more recently than the phylogenetic cutoff.

Further information can be inferred about the time of each HGT or LSE event based on the set of genomes containing the HPK domain (*i.e.*, was the HPK domain

duplicated before or after the divergence of its host species with some other species?).

We have been careful throughout this discussion to refer to the HPK domain rather than the HPK itself – this is to emphasize the fact that we are inferring histories for only the HPK domain of signaling proteins, which (as shown in the Results) can be very different from the evolutionary histories of associated signaling domains that are usually found within the same gene.

Key assumptions

Our approach makes several key assumptions. First, we rely on an accurate species tree. Our species tree compares quite favorably with other published phylogenies [9, 10], and is available as part of the Supplementary Dataset S3. Further, we restrict our analysis to nodes with high bootstrap support. Further, there is no real consensus among researchers as to the topology of the deepest branches, so even well-supported branching patterns according to concatenated gene trees may disagree with trees produced using other methods. We restrict our analysis to relatively recent evolutionary events only, in part to avoid complications that result from uncertainty in these deepest branches. We also make the key assumption that under a model of vertical inheritance, the best BLASTp hit of each domain should be to its most closely-related ortholog. This plausible assumption can be violated for a number of reasons, including unequal evolutionary rates among different lineages. To minimize problems of this type, we impose a stronger cutoff to conclude ‘absence’ of a gene from an outgroup based on a more similar homolog in a distant outgroup: the BLASTp score to the more distant outgroup must be greater than 20 in raw bit score. All of the raw scores used in this analysis are available for browsing online at (<http://microbesonline.org/hpk>). Further,

because we require domains to be ‘absent’ from two or three consecutive well-supported outgroups, and focus only on recent evolutionary events, we feel that it is very unlikely that this assumption is violated in a significant enough number of cases to affect clear trends reported in this study.

Results

An overview of HPKs across bacteria

The fraction of HPKs coded by a given genome is known to scale roughly with the size of the genome, as shown in Figure 2 (an even better correlation is seen if all signaling proteins in a genome are considered [2]). We wanted to identify and investigate genomes that had particularly high numbers of HPKs to see if we could identify their origin. Did these genomes duplicate existing HPKs, acquire large amounts through HGT, or did they simply lose fewer ‘old’ HPKs than other genomes? Several different types of genomes were chosen as examples for more in-depth study throughout this manuscript (and details for all are available in Supplementary Dataset S1). First, we chose organisms in which more than 1.5% of the genome codes for HPKs (red squares). We also targeted two genomes that had the largest numbers of genes acquired by HGT, *Ralstonia solanacearum* and *Pseudomonas syringae* (blue triangles), and one genome in which nearly every new HPK gene was acquired through LSE, *Streptomyces coelicolor* (pink diamond). We chose *Bradyrhizobium japonicum* (cyan diamond) because it includes large numbers of new HPKs acquired through both HGT and LSE, so we could compare these two processes in a single genome. Finally, we included the model

organisms *Bacillus subtilis* and *Escherichia coli*, whose HPKs are the most well-studied experimentally (green circles).

Different species rely on different mechanisms for acquiring new HPKs

Figure 3 summarizes the quantitative results of our phylogenetic analysis across all bacteria (Figure 3A) and individually for each of our targeted genomes (Figure 3B). New HPKs are common across the Bacteria; however, different genomes encode different numbers of new genes. Bacteria in the δ - and ϵ - groups of the Proteobacteria contain particularly high numbers of recently acquired HPKs.

The number of new HPKs arising through HGT or LSE is quite variable across different phylogenetic groups as shown in Figure 4. In some genomes, such as *E. coli* and *Ralstonia solanacearum*, recent gene duplications are rare. HGT, on the other hand, accounts for nearly all of the recently acquired HPKs in these genomes. For others, such as *Desulfovibrio vulgaris* and *Geobacter sulfurreducens*, LSE accounts for the majority of recently acquired HPKs. *Streptomyces* spp. are known for their propensity for gene duplication [11], and their new HPKs result almost exclusively from LSE. The mechanism of gene duplication in *Streptomyces coelicolor* is qualitatively unlike that of other genomes in this study and this point is discussed in greater detail in following sections.

The question of why different genomes have different preferences for HGT or LSE as a means of acquiring new signaling proteins is not obvious, but we did find that genomes with unusually large numbers of HPKs relative to their genome size tend to have accumulated those HPKs *via* LSE. The fraction of HPKs in a genome involved in

recent LSE correlates strongly with the total number of HPKs in that genome (ordinary least squares linear regression: $r=0.74$, $p<10^{-15}$), while the fraction involved in a recent HGT event does not ($r<0.1$, $p=0.93$). In fact, all of the genomes that devote at least 1.5% of their genes to encoding HPKs (*Nostoc* sp. PCC 7120, *Geobacter* spp., *Desulfovibrio* spp., *Wolinella*, and *Dechloromonas*), which are highlighted in red in Figure 2, have major lineage-specific expansions.

In addition, while all of these genomes (again excluding *Nostoc*) are dissimilatory sulfur- or sulfate-reducing bacteria, and many are more closely related δ -proteobacteria, they do not necessarily contain the same expansions. For example, the *Geobacter* lineage contains a large expansion in ‘type 3’ HPKs (using the standard nomenclature defined in [12]), while the *Desulfovibrio* lineage contains an expansion in type 4 HPKs. A further expansion of the ‘hybrid’ type 1b family (including both histidine kinase and response regulator domains in the same protein) is seen only in *Desulfovibrio vulgaris*, and not in its close relative *Desulfovibrio alaskensis* G20 (also known as *Desulfovibrio desulfuricans* G20). Thus, while the propensity for gene duplication may be an inherited trait among these broadly related δ -proteobacteria, the major expansions in each organism are not necessarily shared.

LSE disrupts HPK-RR operon structure

Compared to new HPKs acquired through HGT, HPKs resulting from LSE are less likely to have co-evolved with their cognate response regulators in a single duplication event. Figure 5 shows the distance of new HPKs to response regulators. The data shown do not include ‘hybrid’ HPKs (HPK and RR domains in the same polypeptide chain), which can bias analysis due to their apparent propensity for LSE and since there is

already a RR in the same gene by definition. Averaged over all genomes or taken individually for particular genomes, the trend is clear – LSE genes are much more likely to be present as ‘orphans’, separated from their cognate RRs in the genome. *Streptomyces coelicolor* is an unusual exception to this trend, as it has high numbers of RRs in the immediate proximity of duplicated HPKs. To confirm that operons were the most likely explanation for this genomic proximity between HPKs and RRs, we also compared these different classes of HPKs to operon predictions that have been validated across a wide range of species [13, 14], and observed the same trend: 77% of HGT HPKs had a co-operonic RR, compared to 69% for ‘Old’ HPKs, and only 42% for LSEs.

This separation between HPK and RR evolutionary events suggests that these novel LSEs may be more likely to engage in crosstalk. This is certainly the case for the sole LSE in *B. subtilis*, which is made up of the *kin* regulators of sporulation. *KinA-E* are thought to integrate signals into a common downstream target based on their approximately equal affinity for the regulator *Spo0F* [15]. By contrast, the sole recent duplication in *E. coli*, resulting in the *NarQ/NarX* genes, avoids crosstalk as each HPK ties into a distinct regulator (*NarP* and *NarL* respectively) [16, 17]. A recent study by Laub and coworkers in *Caulobacter crescentus* also found little evidence for physiologically relevant crosstalk among histidine kinases [18]. If crosstalk does not play a large role in general, we would expect to see that the number of ‘orphan’ RRs (not in proximity to a HPK) would generally correlate with the number of ‘orphan’ HPKs (not in proximity to a RR). Figure 6 shows that this trend largely holds across the species examined, though many species show large deviations. We suspect that while some crosstalk may indeed occur, the results from Laub and coworkers are likely to apply to

some extent even across species with large numbers of duplications. Experimental work in these species will be necessary to answer these important questions.

In some cases, we observed that one or a small number of HPKs in an expansion are positioned in operons with RRs. Although beyond the scope of this study, an interesting hypothesis is that these HPKs may be the progenitors of the expansions. For example, *NarQ* is co-operonic with *NarP*, while its duplicate *NarX* is transcribed separately from its cognate regulator, *NarL*. We also observed that the small numbers of HGT genes in genomes with large LSEs are likely to have cognate RRs nearby. This may not only reflect the fact that HPKs are likely to transfer into a genome with their cognate regulators, but also that those HPKs near their cognate regulators make better candidates for transfer out of a genome than their paralogous copies. Indeed, we recently reported a relationship between operons and HGT [5]. We found that nearly 50% of new HGT genes in *E. coli* were acquired with another gene as part of a horizontally transferred operon.

Domain-shuffling often accompanies LSE

HPKs generated by LSE also display more novel variation in their (usually N-terminal) sensory domains than those acquired horizontally. Across all genomes, 47.4% of horizontally transferred HPKs retain a set of upstream signaling domains identical (both in domain type and linear order) to their inferred HGT partner, whereas only 29.1% of recent duplications retain the same domain structure as their closest paralog. In fact, for expansions that include five or more proteins, only 19.9% of closest paralogs had an identical set of upstream domains. Figure 7 shows results for individual genomes. The fraction of HGT genes with conserved upstream domains are shown for those genomes

rich in HGT events, and the fraction of LSE genes with conserved upstream domains are shown for those genomes rich in LSE. For *Bradyrhizobium japonicum*, which contains a mixture of both types of new genes, both numbers are shown. As a control, we also considered a more stringent definition of HGT requiring genes to be absent from three consecutive outgroups. Using this more stringent procedure, 47.3% of horizontally transferred HPKs were found to retain an identical series of upstream signaling domains, which is nearly identical to the 47.4% obtained from the less stringent definition. In addition, we considered the possibility that horizontally transferred HPKs might have a tendency not to include any additional signaling domains, and therefore may be identical trivially. We found that only 10 of our 420 HGT genes lacked any signaling domains, supporting our original conclusions.

These results are particularly striking, since the horizontally transferred HPK domains are on average less similar (lower BLASTp sequence identity) than paralogous domains. These results are also encouraging because our evolutionary inference methods are based only on the similarity of the histidine kinase domain of each HPK, and the high rate of similarity of these upstream signaling domains between putative HGT partners supports the accuracy of our approach. In these results, we considered genes derived from an HGT event followed by a duplication event in the totals for duplicates, but not when computing the totals for HGT as it is not possible using our method to determine which of the resulting paralogs is more likely to have retained the ancestral state of signaling domains.

A notable outlier to Figure 7 is worth mentioning. *Streptomyces coelicolor* contains the largest fraction of new genes acquired by LSE of all the genomes we studied,

yet a large fraction of these genes contain an identical set of upstream signaling domains. In addition, as reported in a previous section, these LSEs tend to involve duplications that preserve HPK-RR pairings. These qualitative differences may reflect an enhanced capability of this genome to duplicate regions of its linear chromosome, a process that has been proposed previously based on genome sequence analysis [11].

Taken together, the results presented in this section suggest different roles for HGT and LSE in HPK evolution. While both mechanisms contribute to the diversity of signaling systems, LSE is accompanied by rearrangements in domain structure as well as by independent evolution of HPK and RR genes. By contrast, other organisms such as *B. subtilis*, *E. coli*, and *Pseudomonas syringae* appear to acquire new HPKs *via* horizontal transfer of intact two-component systems. These consumers of pre-existing genetic diversity are less likely to contain completely novel domain structure, and are more likely to include HPKs in proximity to their cognate RR. Individual genomes appear to have very different preferences for HGT or LSE. LSE is the dominant force in species that are the most highly regulated (those with the highest proportion of genes coding for HPKs), whereas HGT appears to be dominant, for example, in the well-studied model systems *E. coli* and *B. subtilis*.

Anatomy of a lineage-specific expansion

To better understand the structure of a LSE, we investigated a single expansion in the two sequenced *Desulfovibrio* species. Several striking features are present in the expansion depicted in Figure 8. First, the diversity in the upstream signaling domains is obvious in the non-orthologous pairs of HPKs (some likely orthologs between *D. vulgaris* and *D. alaskensis* are shown, and have a similar set of upstream domains).

Second, there was likely a HGT event between *Desulfovibrio* and *Pseudomonas* (probable orthologs from three *Pseudomonas* species are shown in the tree), which likely conserved an upstream domain structure (TM-TM-HAMP-PAS-HPK). This domain structure is identical between the *Pseudomonas* species and one of the members of the *Desulfovibrio* expansion which we postulate served as the donor or acceptor. Third, many of the upstream signaling regions contain repeated domains, but only some of these are noticeably more similar in sequence than other pairs. Thus, rearrangements involve domains that are acquired from distant sources or domains that have been subject to more rapid evolution than histidine kinase domains. Finally, there appears to be a mixture of proteins with and without predicted transmembrane domains, implying that the same basic architecture can support both kinds of signaling mechanisms. Further domain-shuffling may also be happening at the level of the extracellular sensory regions not detected by our sequence profiles.

A close inspection of Figure 8 reveals a pattern in the signaling domain architecture of this expansion: every gene observed has a PAS domain immediately upstream of the histidine kinase domain. Upon closer inspection, we found that this domain is not only conserved in its placement relative to the HPK domain, but is also highly conserved at the sequence level in most of the genes in this family, with clear sequence homology detectable even in the *Pseudomonas* species. This implies that domain architecture is not completely plastic. Instead, there appear to be ‘rules’ for constructing new functional paralogs, and certain domains may be necessary to preserve optimal activity. Similarly, several expansions in *Nostoc* (not shown) consist of a conserved set of core domains preceded by a variable upstream (N-terminal) region.

Some other expansions we studied did not display such obvious patterns of domain architecture. The role of these conserved and non-conserved domains and their mechanism of interaction remains a key open question.

New functional roles for recently duplicated paralogs

LSEs contain a diversity of upstream signaling domains, suggesting that they might respond to different environmental signals. To test this hypothesis, we analyzed microarray data collected for *D. vulgaris* under a variety of stress response conditions to determine whether paralogs had similar expression patterns. Surprisingly, Figure 9 reveals no detectable similarity in gene expression patterns among close paralogs, nor overall similarity within the two *Desulfovibrio*-specific clusters of HPKs. The correlations of gene expression profiles for closest paralogs is not significantly different from those observed between random pairs of genes as measured by the Student's *t* or Kolmogorov-Smirnov tests (as implemented in the R statistical computing package, <http://www.r-project.org/>). As a control, HPKs and their cognate RRs (predicted based on genomic proximity) are strongly correlated within this same data set (see Supplementary Figure S1).

The difference in gene expression patterns and the domain-shuffling both support the idea that these new paralogs have adopted new functional roles within the cell. It is not within the scope of this work to determine the environmental stimuli to which each HPK responds, yet some idea of the variety of possible direct or indirect signals can be inferred from Figure 9. For example, in cluster 1, a paralog (MicrobesOnline accession number 208953, <http://microbesonline.org/>) with domain structure TM-PAS-PAS-PAS-HPK responds strongly to heat shock, and (to a lesser extent) nitrite stress, while a close

paralog (206163) with domain structure TM-TM-HAMP-PAS-HPK responds most strongly to salt stress.

It is important to note that gene expression is an imperfect measure of function. Moreover, many HPKs may be expressed constitutively and regulated mainly at the level of phosphorylation. Nonetheless, we observe some clear cases in which expression is either up- or down-regulated, and those trends are not generally conserved within these phylogenetic clusters. In some sense, signaling genes that are expressed under different sets of conditions could be considered to have different functions even if they regulated overlapping sets of genes. We feel that the general lack of co-expression, when combined with the diversity of newly evolved signaling domain architectures, together make a strong case for new functional roles.

Genomic distribution of HPK families

We looked at the distribution of paralogs within each genome to see if we could infer any information regarding the mechanism of gene family expansion. In *B. subtilis*, for example, all five of the *kin* genes are contained within a small region of the chromosome, with four of them very tightly spaced (the LSE or purple-colored genes in Figure 10). In general, however, we observed very little clustering of genes within genomes. To be more rigorous, we constructed a simple statistical test to measure clustering of new HPKs in a genome. We computed the distribution of nearest-neighbor distances between HPKs arising from LSE, and compared this to the distribution expected by chance (approximated by an exponential distribution with mean=[# of genes in genome]/[# of recent LSEs]). We then used the Kolmogorov-Smirnov test to determine if the two distributions were significantly different. Of the genomes we

classified as having large numbers of LSEs, only *Nostoc* showed significant clustering. When we examined this result further, we identified the source of the clustering: a set of two adjacent HPKs, which likely work together to relay signals. The first gene in each of these pairs contains a wide variety of largely shuffled signaling domains, while the downstream gene contains a conserved HPK domain followed by a CheY-type regulator domain. After correcting for this by counting these adjacent pairs as a single duplication, we observed no clustering among LSE genes in *Nostoc*. Figure 10 shows an overview of genomic positions of HPKs and RRs across several species, none of which (apart from the *kin* locus of *B. subtilis*) appear to have significant clustering. Thus, the duplication of HPKs appears qualitatively different from the duplication of signaling domains within the N-terminal region of individual HPKs, as the latter often occur in long tandem stretches.

Timing of evolutionary events

Because our phylogenetic inference procedure identifies LSE and HGT events back to a particular outgroup, we can trace the influx of HPKs into each lineage as a function of time. Figure 11 shows the number of HPKs predicted to have entered several lineages as a function of time. While different species here show different overall trends (some like *Pseudomonas syringae* gradually accumulated HPKs, while some like *Nostoc* acquired most of their HPKs very recently), the species-averaged plot shows a steady influx of HPKs at a nearly constant rate back until about our phylogenetic cutoff distance of 1.0 (where HGT tends to saturate since it must be missing from at least two outgroups predating the transfer). Moreover, both HGT and LSE seem to be contributing at similar levels to the total number of HPKs, and both accumulate at about the same rate. It is important too note that the resolution of these figures depends directly on the number of

sequenced bacterial groups at different levels of divergence from each genome, and caution should be used when comparing our evolutionary distances across distant taxa as differences in evolutionary rate were not explicitly considered in this analysis. As more genome sequences become available, it will be possible to resolve the timing of these events with higher resolution, and even measure turnover rates for HPKs.

Discussion

Different strategies for obtaining new genetic diversity

Histidine protein kinases (HPKs) play a key role in allowing bacterial cells to sense and respond to their environment. We investigated specifically those HPKs that evolved recently they are the most likely to shed light on how bacteria adapt to their particular niches. We investigated horizontal gene transfer and lineage-specific gene family expansion of HPKs. Among those genomes most likely to be highly regulated *via* environmental signals (*i.e.*, those with the largest fraction of genes coding for HPKs), recently acquired genes came mostly from gene family expansion. Moreover, we found that gene family expansion is often accompanied by domain shuffling to produce signaling proteins with unique combinations of sensory domains, and that gene expression patterns among these paralogous genes were highly diverged. Together, these results suggest that gene family expansion and domain shuffling serve to facilitate the repurposing of existing signaling proteins to new tasks, but only in a subset of species. These species have adapted to their environment by trying new combinations of existing signaling domains to generate more novel genetic diversity. By contrast, the well-studied model organisms *E. coli* and *B. subtilis* likely acquire most of their new HPKs from

existing bacteria, intact with identical sensory domains in many cases and together in operons with cognate response regulators.

HGT and LSE across the bacteria

Several recent studies [6, 19-21] have begun to add detail to a framework for understanding the evolution of gene content in bacterial genomes. A key finding of the survey of gene acquisition/loss by Lerat and coworkers [19] is that gene family expansion is not a major mechanism of gene acquisition, at least in γ -Proteobacteria. The current work focuses on a notable exception to this general rule, the histidine protein kinases. It is worth noting that in the set of genomes examined by Lerat and coworkers, HPKs did follow the general trend observed for other genes, new acquisitions generally represented an influx of new gene families rather than duplication of existing families. Therefore, it may be premature to conclude that it is this family of genes rather than the genomes themselves that are exceptional in this regard. Very large expansions of cytochrome genes have been reported for some of the δ -proteobacteria that we follow in this study [22], lending some support to the idea of genome-specific in addition to gene-specific differences in the propensity for expansion.

Whether or not some genomes have greater propensities for gene duplication overall, there are genome-specific differences in the likelihood of duplications in the HPK gene family. Expansion of HPKs is enriched in a subset of bacteria, primarily in early-branching Proteobacteria. In contrast, the model organisms *E. coli* and *B. subtilis* rely largely on a set of ancient HPKs and some recently acquired *via* horizontal gene transfer. As a result, the former group of environmental bacteria contains a larger

number of total HPKs, and their HPKs tend to have more unique combinations of upstream signaling domains.

Alternative explanations for HGT

LSE events are more straightforward to interpret than HGT events in our phylogenetic framework. Few would argue with the assertion that a gene present in multiple highly similar copies in a single clade, but only one (or zero) more distantly-related copies in earlier-branching outgroups reflects a lineage-specific expansion. Identification of HGT, on the other hand, relies on a less obvious argument: that a single HGT event is a more parsimonious (and therefore more likely) explanation for the absence of a gene in two or more consecutive outgroups (excluding reduced genomes) than is two or more consecutive deletion events. Therefore, some additional consideration of the evidence for HGT and alternative explanations for the observed patterns of gene distributions is called for.

One of the key lines of evidence supporting our approach for identifying HGT events is that we are able to identify donor-acceptor pairs with identical upstream signaling domains using only sequence data from the HPK domain. While this lends support to our domain-based BLASTp approach, it does not directly address the challenge of identifying a HGT scenario from multiple deletions, since vertically transmitted orthologs would also be likely to retain a similar set of upstream signaling domains. To test this hypothesis, we used a more conservative cutoff, absence from three consecutive outgroups, for identifying HGT events. This stricter cutoff produced nearly identical results (47.3% versus 47.4% of upstream domains were conserved using the two definitions), as reported in the earlier section on domain shuffling, lending support to a

true difference between HGT and LSE genes. In addition, we found that horizontally transferred HPKs were *more* likely than vertically-transmitted genes to occur in the same operon as a response regulators, supporting a distinction between these HGT genes and vertically transmitted genes.

We also examined the distribution of donor/acceptor genome pairs for the genomes with large amounts of inferred HGT events. While there was a general trend for HGT events within major bacterial groups (*e.g.*, between *B. subtilis* and other Firmicutes, or between *E. coli* and other Proteobacteria), there was not a clear excess of transfer to and from a single pair of species, which might be expected if there were significant false positives resulting from an inaccurate species tree.

Finally, we observe that our results are in general agreement with a recently published study that estimated HGT rates across all gene families among sequenced bacteria [20]. *Bradyrhizobium japonicum* has one of the highest rates of HGT among all bacteria in both studies, and most of the genomes indicated as having high rates of HGT in the previous study indeed are indicated to have significant levels of new HPKs arising through HGT in the present work. It should be noted that we did not describe results for *Pirellula* sp. in the current study because we could not resolve its phylogeny well enough to make confident assertions about its evolutionary history. Another difference between our study and the work by Ouzounis and coworkers [20], is our finding that the sequenced *Pseudomonas* species are outliers containing a large number of HPKs possibly acquired *via* HGT. *Pseudomonas* was also identified by [19] as having recently acquired a large number of ‘ORFan’ genes without obvious sequence homologs, but we did not see evidence for large numbers of novel HPK ‘birth’ events. Whether *Pseudomonas*

genomes confine their horizontal transfer to genes involved in signal transduction is an interesting question for further study.

Mechanisms of HPK evolution

The mechanisms that can lead to HGT of genes are well studied. Phage, plasmids, and competence can all help facilitate the transfer of foreign DNA. In a dramatic example, most of the HPK genes (10/17) on the megaplasmid of *Ralstonia solanacearum* are predicted to be horizontally transferred into that lineage, and 9 out of 10 of these are inferred to have transferred very recently (*i.e.*, they are absent in the closest two outgroups). This suggests the megaplasmid may have served as a vessel for the import of these genes (see plasmid genes in Figure 10). In contrast, only 11 of 27 genes on the main chromosome are inferred to be the result of horizontal transfer, and of these, only 5 are absent in the closest outgroups.

Despite the fact that LSE is a major driving force in the genesis of new HPKs (by our estimates, a larger factor than HGT), the mechanism(s) behind these expansions are unclear. In some cases, transposons can be implicated in recent gene expansions, in particular, when the transposase genes occur next to the gene copies undergoing expansion. Yet, for most of the large expansions we observed, there was no evidence of nearby transposons and duplications of HPKs were accompanied by domain shuffling of their upstream regions. Transposons are unsatisfying as an explanation in this case because new combinations of upstream signaling domains are not created directly from single transposition events (an insertion event would interrupt the reading frame between the HPK domain and the new signaling domain). Thus, genomes involving large LSEs may contain additional mechanisms that allow for the rearrangement of individual

domains, and the presence of these mechanisms may account for the predisposition of those genomes to LSE rather than HGT.

A recent genomic survey of recombination machinery across bacteria did not identify specific components unique to this set of genomes [23]. Instead, many of the species we identified as rich in gene expansions were highlighted as genomes known to undergo recombination, but lacking a full complement of presynaptic recombination proteins. Perhaps some of these genomes contain alternative, yet to be identified presynaptic genes with properties that facilitate LSE and/or domain-shuffling. Gene duplication is known to happen at high frequency near the ends of the linear *Streptomyces* chromosome [11, 24], but probably through a different mechanism than other LSE-rich genomes because the domain structure and cognate response regulators are more highly conserved in this species compared to the other species prone to duplication events.

An alternative to the theory of genome-specific recombinational machinery is a role for phage or other extrachromosomal elements in the duplication and rearrangement of domains in these signaling proteins. A recent study by Aravind and coworkers found massive domain shuffling and LSE of a family of fungal transcription factors occurring inside viral genomes [25]. A phage origin for LSEs would also explain the discrepancy between the relatively common tandem duplication and shuffling of signaling domains within HPKs (which could occur in the phage genome), and the apparent random positioning of paralogous HPKs across the genome (which would be the result of random phage insertions). To test the phage origin hypothesis, we examined large LSEs across a number of genomes, but did not find a statistically significant excess of phage genes near

paralogs. Moreover, a BLASTp search of phage databases yielded only a few very weak hits to HPKs.

Whether genome-specific recombinational machinery, extrachromosomal elements, or other factors (such as environments that select for large numbers of HPKs) are responsible for generating genetic diversity among HPKs, it is clear that there are genome-specific differences. Genomes such as *Streptomyces* represent still a different paradigm (*i.e.*, capability for large-scale expansion but little new diversity, at least at the level of domain structure). Why this creative ability appears to be limited to a subset of genomes, and whether these differences among taxa extend to other gene families, emerge as key directions for future research.

Methods

Inferring evolutionary histories of HPKs.

To describe the origin of HPK domains, we examined the phylogenetic distribution of those domains across other taxa. Similar to previous work [5, 6, 20, 26], we inferred ‘birth’ events when a domain was found only in a single clade and horizontal transfer events in cases where multiple gene losses were needed to explain the phylogenetic distribution assuming vertical descent. Unlike the previous studies mentioned, we were interested in cataloging and determining the relative age of gene duplication events. Because different HPK families can be considered ancient paralogs, we were not able to use a simple presence/absence criterion to describe the distribution of genes in taxa. Instead, we used a BLASTp score cutoff on sequence similarities to infer the phylogenetic distributions of particular subfamilies as described in the following

section. Importantly, our analysis was based only on the histidine kinase domains (on average 224 residues) excluding sensory and other domains, so that we could discriminate the independent processes of domain evolution and shuffling of upstream signaling domains.

The inference process is briefly outlined below, and details on each step are provided in the following sections.

Construct phylogeny for all species using concatenated ubiquitous genes

Identify HPKs from coding regions of all genomes

Parse HPKs into domains (PAS, HAMP, TM, HPK, etc.)

Find best hits of each 'query' HPK (domain) to outgroups:

Move up one node in the species tree (starting from the genome containing the 'query' HPK)

Add the leaves of the new branch to the current 'outgroup'

If the current node has a bootstrap value of less than 80% then repeat step 4a

Record the best hit of the 'query' HPK to the current 'outgroup' species

Record the number of 'paralog' HPKs (in the same genome as the query) that have higher BLASTp scores than the best hit to the current outgroup

The gene count at this node (see step 6) is equal to $(1/\# \text{ of paralogs})$

The duplicate count at this node (see step 6) is equal to $[1 - (\text{the gene count})]$

Repeat step 4a-e with a new 'outgroup' until the root node is reached

Build presence/absence profile from best hits:

If the best hit to any outgroup is less than 25% of the BLASTp score of the query HPK to itself, then set that outgroup to zero (for absent)

Else, starting from the most ancient outgroup (*Aquifex aeolicus*), set the score of each outgroup equal to one (for present) unless the best hit to that outgroup is less than the best hit to an older outgroup by a bitscore value of 20; otherwise set the score to zero (for absent)

If any outgroup with score zero contains only reduced genomes with fewer than 10 HPKs, change its score to two (for unknown)

Infer evolutionary history from presence/absence profile:

If there are two or more consecutive zeroes (ignoring twos) in the presence/absence profile for a given HPK, then it is considered HGT

The oldest outgroup containing a one before the run of zeroes is considered the age of the HGT event

If this age is past the phylogenetic cutoff, then the HGT is not counted

Else, the gene count (step 4.e.i) at that node is added to the number of HGT events for the query genome

Else, if there are no ones in outgroups older than the phylogenetic cutoff distance, then it is considered a HPK subfamily 'birth', and the gene event score at the oldest 'one' is added to the estimated number of 'birth' events.

The number of gene duplications is estimated by adding the duplicate count (step 4.e.ii) at the oldest outgroup containing a ‘one’ that is within the phylogenetic cutoff age to the total LSE counts.

1. Construction of bacterial species tree.

Each gene that was present in every bacterial genome studied, without obvious paralogs (no other genes in the same COG), was used to construct a species tree. The tree was built based on the multiple sequence alignment (MSA) using the concatenated sequences of these 15 genes, which are listed in the Supplementary Table S1. Muscle [27] was used for the MSA; gaps were trimmed using MEGA3 [28]. A neighbor-joining tree with 100 bootstrap replicates was built using the Phylip [29] software package with the PMB (Probability Matrix from Blocks) [30] amino acid substitution model. Gamma-distributed rates were used with a shape parameter of 0.72, which was estimated using TREE-PUZZLE [31]. Branch lengths were then constructed using the bootstrap consensus tree topology and the PROTDIST program included in the PHYLIP package [29] enforcing equal distances of each species to the tree root. Unequal rates for different taxa were not taken into consideration, and therefore distances should not be interpreted as accurate estimates of evolutionary time. The tree is included in the Supplementary Dataset S3 for review.

Our species tree is based on concatenated protein (mostly ribosomal) genes. Similar approaches have been shown to be effective at producing well-resolved trees [9, 10]. We also experimented with 16S rDNA sequence trees and genome content trees, but found that the concatenated gene trees produced better-resolved trees that seemed to avoid some unexpected groupings that probably arise from differences in GC-content and

reduced genomes. We include our species tree in Newick format as Supplementary Dataset S3 for review. We chose the root of our tree at the last common ancestor of *Aquifex aeolicus* and other bacteria. We do not claim strong evidence for the early branching of this lineage, instead we limit our analysis to more recent evolutionary events that do not depend strongly on the topology of the deepest-branching nodes.

2. Identification of HPKs.

We identified 4,959 HPKs in 20 archaeal and 187 bacterial genomes based on sequence and profile similarities. A protein was considered a putative HPK if the protein contained a histidine kinase domain, (IPR005467 as measured using the InterPro software suite [32]), or was assigned to the signal transduction histidine kinase COG4582 using a profile-based RPS-BLAST search [33, 34]. Additionally, we considered all proteins that were not picked up by either method, but contained an ATPase domain according to SuperFamily motif SSF55874 [35]. If a protein contained an ATPase domain and was a transmembrane protein or had one or more known signaling domains (*e.g.* PAS, HAMP), then it was also considered a HPK.

For 123 out of 130 genomes, our approach compared favorably to that of [2] which used manually-curated PSI-BLAST searches to identify likely HPKs, in that we identified nearly the same number of HPKs. For seven genomes, the total counts differed by more than 4, with *Streptomyces coelicolor* being the largest deviation (the study by Galperin identified 95, while our method found only 81). Overall, we argue that our main findings are not substantially affected by the differences in the numbers of HPKs as our overall numbers are consistent with the manually curated study by Galperin. A detailed comparison on a genome-by-genome basis is provided as Supplementary Dataset

S1. A FASTA format file including all HPK domains used in this analysis is included as Supplementary Dataset S4.

No HPK domains were found in sequenced genomes within the taxonomic group Mollicutes. Bacterial genomes with small genome size (i.e. <1000 protein coding genes) tended to have few or no HPKs, and the proportion of HPKs was correlated with the genome size. 11 out of 20 archaeal genomes with a large range of genome sizes (553-3106 protein coding genes) had no HPKs. HPKs from archaeal genomes were not considered in this analysis.

3. Identification of structural and signaling domains in histidine kinases.

We used the profile-based methods included in the InterPro software suite [32] to identify common signaling domains, and TmHMM [36] to predict transmembrane regions of HPKs. The key protein families used were: CheY-like response regulator receiver: PF00072, SSF52172; CheB methylesterase: PS50122; CheR-type MCP methyltransferase: PS50123; PAS: SM00091, SSF55785, or TIGR00229; HAMP: PF00672, PS50885; GAF: SSF55781, SM00065; Hpt: SSF47226, PF01627; CACHE: PF02743; Phytochrome, light-sensing: PF00360, PS50046 (PF=Pfam, SM=SMART, SSF=SuperFamily, TIGR=TIGRFAM, PS=PROSITE PROFILE).

4. Build best BLASTp hit profile.

The first step in this procedure was to identify a set of unambiguous ‘outgroups’ for each species. We considered each ancestor node on the tree built in step one that contained the target species, starting with the most recent. If the bootstrap support at that node was at least 80%, then all the species present in the leaves of the new branch were

considered to constitute an ‘outgroup’. If the bootstrap support was less than 80%, then those species were put aside until an internal node with at least 80% support was reached, and all of the leaves of those new branches were combined into a single ‘outgroup’. The best BLASTp hit to any of the HPK domains contained within each outgroup was recorded.

5. Build presence/absence profile from best hits.

Each best hit profile was converted to string of integers (0, 1, or 2), indicating whether the particular gene subfamily was likely to be present or absent from each outgroup. First, a lower bound was placed on BLASTp hits: scores less than 25% of the maximal BLASTp score (of the query gene to itself) were not considered, as these are essentially different subfamilies of HPKs. These low scores were set to zero.

Outgroups that had BLASTp hits greater than the best hit to any more distant outgroups (minus a threshold bitscore of 20 from the distant outgroup hit), were assigned a one. A string of ones, therefore, indicates a set of BLASTp distances consistent with vertical inheritance of the query HPK. The threshold of 20 makes our method more conservative in calling gene absences, and was intended to lessen the effects of evolutionary rate differences among lineages and possible errors in our species tree topology.

Outgroups without a hit greater than that seen in older groups were assigned zero, unless every genome in the outgroup had less than 10 HPKs. In that case, it is likely that those species had undergone a genome-wide reduction in the number of HPKs, and multiple absences from such outgroups does not provide strong support for the alternative hypothesis of HGT. These outgroups were assigned a value of 2. The cutoff of 10 HPKs

is essentially an *ad hoc* rule that works well at identifying reduced genomes in the set of genomes studied. A list of excluded genomes is given in Supplementary Dataset S2. A website including all the raw BLASTp scores, their presence/absence profiles, and other key information for each HPK in this study is provided at <http://microbesonline.org/hpk>.

6. Inferring evolutionary events from presence/absence profiles.

We sought to identify the events that led to the complement of HPKs observed in each genome contained in the MicrobesOnline database [37] as of February 2005. We considered four possible origins for extant HPKs: (i) duplication of pre-existing HPKs, (ii) horizontal transfer from distantly related genomes, (iii) ‘birth’ of novel HPK subfamilies, and (iv) ‘old’ HPKs that were present in the genome early in its evolutionary history.

Phylogenetic cutoff. To classify evolutionary events as ‘recent’, we defined a phylogenetic cutoff distance before which we did not report events. This had the additional benefit of making our analysis robust to the topology of the deepest and most difficult to resolve branches. The cutoff distance used is roughly equivalent to the divergence time of *E. coli* from the most distantly related γ -proteobacteria. For details on the cutoff for different lineages see the tree and the website provided at (<http://microbesonline.org/hpk>). Because different taxa have different evolutionary rates for the same set of genes, we used a cutoff based on a species tree in which the distance of each leaf to the root is assumed to be equal. We used this ‘linearized’ tree distance not to compute accurate divergence times, but to enforce our phylogenetic cutoff distance across different lineages more evenly.

‘Event’ counting. The basic algorithm for counting HGT, LSE and Birth events is given in the outline above, but some explanation is necessary for the ‘gene count’ and ‘duplicate count’ calculations. Because we consider each HPK independently, we need to adjust the counts such that a gene that is horizontally transferred once, and subsequently duplicated is not counted as two HGT events. For this reason, we count only $(1/\# \text{ of paralogs})$ HGT events for each duplicate copy. By similar reasoning, a gene duplicated four times should not be counted as five duplication events, because one copy represents the original gene. To avoid overcounting of duplication events, we count only $[1 - (1/\# \text{ of paralogs})]$ LSE events for each paralog.

Gene expression data. Gene expression microarray data was downloaded from the MicrobesOnline database [37]. A full description of the salt, heat, and nitrite stress experiments is given in [38-40]. Briefly, all experiments were performed in LS4D (lactate-sulfate) medium under anaerobic conditions. Cells were grown to log phase and then subject to stressors: 8C cold shock (30C control), 50C heat shock (37C control), 1000ppm oxygen, 500mM NaCl, 2.5mM nitrite. Each sample was measured relative to a genomic DNA control, and reported values are the log-ratio of expression levels at the indicated timepoint versus the pre-stress (0 min) levels for the same biological sample.

Supporting Information

Dataset S1.

Detailed raw and processed data for each genome

Found at DOI: [10.1371/journal.pcbi.0020143.sd001](https://doi.org/10.1371/journal.pcbi.0020143.sd001) (138 KB XLS).

Dataset S2.

Genomes excluded from HGT calculation

Found at DOI: 10.1371/journal.pcbi.0020143.sd002 (12 KB XLS).

Dataset S3.

Species tree with distances and bootstrap scores

Found at DOI: 10.1371/journal.pcbi.0020143.sd003 (4 KB TXT).

Dataset S4.

FASTA format file including all HPK domain sequences used

Found at DOI: 10.1371/journal.pcbi.0020143.sd004 (3.0 MB TXT).

Figure S1.

Correlation in expression between operons and response regulators

Found at DOI: 10.1371/journal.pcbi.0020143.sg001 (30 KB DOC).

Table S1.

List of COGs used to build species tree

Found at DOI: 10.1371/journal.pcbi.0020143.st001 (21 KB DOC).

Acknowledgments

We thank Chris Rao, Morgan Price and other members of the Arkin group for insightful discussion and critical evaluations of the manuscript. We thank three anonymous reviewers for suggestions that improved the quality of this manuscript.

Author contributions. EA, KH, and AA conceived and designed the experiments. KH performed the experiments. EA and KH analyzed the data. EA wrote the paper.

Funding. This work was part of the Virtual Institute for Microbial Stress and Survival (<http://VIMSS.lbl.gov>) supported by the U. S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics Program:GTL through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U. S. Department of Energy.
Competing interests. The authors have declared that no competing interests exist.

References

1. Ulrich, L.E., E.V. Koonin, and I.B. Zhulin, One-component systems dominate signal transduction in prokaryotes. *Trends Microbiol*, 2005. 13(2): p. 52-6.
2. Galperin, M.Y., A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. *BMC Microbiol*, 2005. 5: p. 35.
3. Wolanin, P.M., P.A. Thomason, and J.B. Stock, Histidine protein kinases: key signal transducers outside the animal kingdom. *Genome Biol*, 2002. 3(10): p. REVIEWS3013.
4. Rodrigue, A., et al., Two-component systems in *Pseudomonas aeruginosa*: why so many? *Trends Microbiol*, 2000. 8(11): p. 498-504.
5. Price, M.N., et al., Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res.*, 2005. 15: p. 809-19.
6. Daubin, V. and H. Ochman, Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res.*, 2004. 14: p. 1036-42.
7. Ragan, M.A. and R.L. Charlebois, Distributional profiles of homologous open reading frames among bacterial phyla: implications for vertical and lateral transmission. *Int. J. Syst. Evol. Microbiol.*, 2002. 52: p. 777-87.
8. Price, M.N., A.P. Arkin, and E.J. Alm, The life-cycle of operons. *PLoS Genet*, 2006. 2(6): p. e96.

9. Lerat, E., V. Daubin, and N.A. Moran, From Gene Trees to Organismal Phylogeny in Prokaryotes: The Case of the gamma-Proteobacteria. PLoS Biol., 2003. 1: p. E19.
10. Ciccarelli, F.D., et al., Toward automatic reconstruction of a highly resolved tree of life. Science, 2006. 311(5765): p. 1283-7.
11. Bentley, S.D., et al., Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature, 2002. 417(6885): p. 141-7.
12. Grebe, T.W. and J.B. Stock, The histidine protein kinase superfamily. Adv Microb Physiol, 1999. 41: p. 139-227.
13. Price, M.N., et al., A Novel Method for Accurate Operon Predictions in All Sequenced Prokaryotes. Nucleic Acids Res., 2005. 33: p. 880-92.
14. Price, M.N., A.P. Arkin, and E.J. Alm, OpWise: operons aid the identification of differentially expressed genes in bacterial microarray experiments. BMC Bioinformatics, 2006. 7: p. 19.
15. Jiang, M., et al., Multiple histidine kinases regulate entry into stationary phase and sporulation in *Bacillus subtilis*. Mol Microbiol, 2000. 38(3): p. 535-42.
16. Yamamoto, K., et al., Functional characterization in vitro of all two-component signal transduction systems from *Escherichia coli*. J Biol Chem, 2005. 280(2): p. 1448-56.
17. Rabin, R.S. and V. Stewart, Dual response regulators (NarL and NarP) interact with dual sensors (NarX and NarQ) to control nitrate- and nitrite-regulated gene expression in *Escherichia coli* K-12. J Bacteriol, 1993. 175(11): p. 3259-68.

18. Skerker, J.M., et al., Two-component signal transduction pathways regulating growth and cell cycle progression in a bacterium: a system-level analysis. *PLoS Biol*, 2005. 3(10): p. e334.
19. Lerat, E., et al., Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol*, 2005. 3(5): p. e130.
20. Kunin, V., et al., The net of life: reconstructing the microbial phylogenetic network. *Genome Res*, 2005. 15(7): p. 954-9.
21. Snel, B., P. Bork, and M.A. Huynen, Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res*, 2002. 12(1): p. 17-25.
22. Methe, B.A., et al., Genome of *Geobacter sulfurreducens*: metal reduction in subsurface environments. *Science*, 2003. 302(5652): p. 1967-9.
23. Rocha, E.P., E. Cornet, and B. Michel, Comparative and evolutionary analysis of the bacterial homologous recombination systems. *PLoS Genet*, 2005. 1(2): p. e15.
24. Chen, C.W., et al., Once the circle has been broken: dynamics and evolution of *Streptomyces* chromosomes. *Trends Genet*, 2002. 18(10): p. 522-9.
25. Iyer, L.M., E.V. Koonin, and L. Aravind, Extensive domain shuffling in transcription regulators of DNA viruses and implications for the origin of fungal APSES transcription factors. *Genome Biol*, 2002. 3(3): p. RESEARCH0012.
26. Kunin, V. and C.A. Ouzounis, The balance of driving forces during genome evolution in prokaryotes. *Genome Res*, 2003. 13(7): p. 1589-94.
27. Edgar, R.C., MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 2004. 5: p. 113.

28. Kumar, S., K. Tamura, and M. Nei, MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform*, 2004. 5(2): p. 150-63.
29. Felsenstein, J., PHYLIP (Phylogeny Inference Package) version 3.6. 2005, Distributed by the author.
30. Veerassamy, S., A. Smith, and E.R. Tillier, A transition probability model for amino acid substitutions from blocks. *J Comput Biol*, 2003. 10(6): p. 997-1010.
31. Schmidt, H.A., et al., TREE-PUZZLE: Maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics*, 2002. 18: p. 502-504.
32. Quevillon, E., et al., InterProScan: protein domains identifier. *Nucleic Acids Res*, 2005. 33(Web Server issue): p. W116-20.
33. Tatusov, R.L., et al., The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, 2001. 29: p. 22-8.
34. Marchler-Bauer, A., et al., CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res*, 2003. 31(1): p. 383-7.
35. Madera, M., et al., The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res*, 2004. 32(Database issue): p. D235-9.
36. Kall, L., A. Krogh, and E.L. Sonnhammer, A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*, 2004. 338(5): p. 1027-36.
37. Alm, E.J., et al., The MicrobesOnline Web site for comparative genomics. *Genome Res.*, 2005. 15: p. 1015-22.

38. Chhabra, S.R., et al., Global analysis of heat shock response in *Desulfovibrio vulgaris* Hildenborough. *J Bacteriol*, 2006. 188(5): p. 1817-28.
39. He, Q., et al., Energetic consequences of nitrite stress in *Desulfovibrio vulgaris* Hildenborough, inferred from global transcriptional analysis. *Appl Environ Microbiol*, 2006. 72(6): p. 4370-81.
40. Mukhopadhyay, A., et al., Salt stress in *Desulfovibrio vulgaris* Hildenborough: an integrated genomics approach. *J Bacteriol*, 2006. 188(11): p. 4068-78.

Figure 1. Overview of approach

An overview of our phylogenetic inference procedure is given. We look only at histidine kinase domains from HPKs, and compare the distribution of these to the species tree.

When homologs in distant outgroups are more distantly related, we infer simple vertical descent. Paralogs and distances that contradict species phylogeny result in our inference of gene duplication or horizontal transfer. Only events (such as duplication or transfer) that occurred more recently than the cutoff as described in the Methods are considered. Four hypothetical cases are shown, and each is labeled as present/absent ('1' or '0') from each outgroup according to the procedure described in the Methods.

DOI: 10.1371/journal.pcbi.0020143.g001

Figure 2. HPK content *versus* genome size

The percent of each genome is plotted as a function of genome size. As reported in previous studies, there is a roughly linear correlation. Highlighted in colored symbols are several groups of genomes described in the text: genomes coding a high ($\geq 1.5\%$) fraction of HPKs – red squares; the model organisms, *E. coli* and *B. subtilis* – green

circles; genomes with a high number of HGT events – blue triangles; *Bradyrhizobium japonicum* (high number of both HGT and LSE genes) – pink diamond; and *Streptomyces coelicolor* (high percentage of LSE genes) – cyan diamond.

DOI: 10.1371/journal.pcbi.0020143.g002

Figure 3. Summary of evolutionary events

The number of events inferred for different bacteria is summarized in this figure. (A) Average numbers for the major taxonomic groups used in this study. (B) Specific numbers for targeted genomes (those with colored symbols in Figure 1).

DOI: 10.1371/journal.pcbi.0020143.g003

Figure 4. LSE events *versus* HGT events

The number of LSE and HGT events for each genome are shown. Colored symbols correspond to the genomes identified in Figure 2. Note the position of the red squares well above the x-axis.

DOI: 10.1371/journal.pcbi.0020143.g004

Figure 5. Proximity of different classes of HPKs to response regulators

Shown is the cumulative percentage of HPKs of each type that have response regulators within the distance on the chromosome specified by the x-axis. The different gene types shown are: old HPKs – black circles (Old); HGT genes without recent paralogs – blue triangles (HGT); and HPKs with recent paralogs – red crosses (LSE). In the bottom right panel, an average over all genomes is shown. In general, and for most specific cases (excepting *Streptomyces*), horizontally transferred genes are observed to have a much

higher fraction of RRs in close genomic proximity. ‘Hybrid’ HPKs, which have RRs in the same ORF as the HPK, were excluded from this analysis. Only genes that are not believed to have undergone duplication within a lineage are used in the HGT group. Lines stop when cumulative percentage equals 100%.

DOI: 10.1371/journal.pcbi.0020143.g005

Figure 6. Co-evolution of orphan HPKs and RRs

The number of ‘orphan’ HPKs is plotted versus the number of ‘orphan’ RRs. A moderate but highly significant linear correlation is observed (ordinary least squares linear regression: $r=0.57$, $p<10^{-15}$).

DOI: 10.1371/journal.pcbi.0020143.g006

Figure 7. Extent of domain shuffling in different classes of HPKs

The fraction of HPKs with identical upstream domains to either their inferred HGT partners (red bars), or to their closest paralog (blue bars) in the case of LSE. Only genes that are not believed to have undergone duplication within a lineage are used in the HGT group. *Bradyrhizobium japonicum*, which has a significant number of genes classified as HGT and LSE, is shown twice.

DOI: 10.1371/journal.pcbi.0020143.g007

Figure 8. Domain shuffling in a *Desulfovibrio* spp. Expansion

The domain structure of genes in a large LSE in the *Desulfovibrio* genus is shown. In addition, three similar proteins identified in *Pseudomonas* species are shown, which are likely the result of HGT. Genes are identified by their species name and their accession

number in the MicrobesOnline database (<http://microbesonline.org/>) for easy reference (DV refers to genes present in *Desulfovibrio vulgaris* and DA refers to genes present in *Desulfovibrio alaskensis* G20). Each domain corresponds to a branch of the TREE-PUZZLE phylogenetic tree (of only the HPK domains) shown at left. Each PAS domain is colored according to sequence homology (as inferred by BLASTp), and domains with the same color comprise subfamilies of closely related domains. While upstream domains are generally shuffled, each gene shown contains a PAS domain immediately preceding the conserved histidine kinase (HPK) domain. Moreover, this PAS domain is largely conserved among paralogs at the sequence level, while more N-terminal domains are not. Interestingly, the *Pseudomonas* gene, which we infer to be involved in a horizontal transfer event, has a set of signaling domains identical to one of the *Desulfovibrio* copies suggesting a likely donor/acceptor pair, and highlighting the qualitative difference in genes acquired by HGT and LSE.

DOI: 10.1371/journal.pcbi.0020143.g008

Figure 9. Gene expression of *Desulfovibrio vulgaris* expansions

Gene expression profiles across a compendium of experimental stress response conditions (NaCl, heat shock, cold shock, nitrite, and oxygen) were monitored using DNA microarrays, and shown next to a phylogenetic NJ tree (with 1000 bootstraps, generated using the MEGA3 software package [28]) of all HPK domains in *D. vulgaris*. Blue colors indicate down-regulated genes (relative to unperturbed cells), and red colors indicate up-regulated genes. No significant excess correlation in gene expression was observed for genes within each cluster (compared to randomly chosen pairs of genes) using a Student's t-test to compare mean correlations or the Kolmogorov-Smirnov test to

compare distributions. The domain structure of each gene in the three LSEs is shown to the right. Gene names are provided for all genes, and MicrobesOnline accession numbers are provided in parentheses for genes in each of the major clusters for comparison with Figure 8. Bootstrap values are provided for each of the major clusters, and the amino acid sequence of the ‘H-Box’ motif for genes in each cluster is shown. A more detailed description of the experiments performed is given in the Methods.

DOI: 10.1371/journal.pcbi.0020143.g009

Figure 10. Genomic distribution of HPKs and RRs

The position of signaling proteins in several genomes is shown. In the outer ring, HPKs of different classes are shown: Old (gray), LSE (purple), HGT without duplication (blue), and genes that recently underwent a ‘Birth’ event (green). The middle ring shows the position of response regulators with blue colors indicating hybrid response regulators (containing HPK domains). The inner ring shows the location of all genes in each genome annotated as signaling proteins according to the MicrobesOnline database [37].

DOI: 10.1371/journal.pcbi.0020143.g010

Figure 11. HPK evolution over time

The number of HPKs entering a lineage is shown as a function of the time each HPK entered that lineage (*i.e.*, the distance of that species to the last ancestor predicted to contain that HPK). Red lines/diamonds indicate LSE events, and blue lines/triangles indicate HGT events. Plots are cumulative showing all events dating more recently than the time shown on the x-axis. HGT lines do not extend as far to the right as LSE lines, since they require genes to be lost from two consecutive outgroups. The vertical dashed

line shows the phylogenetic cutoff distance, and the horizontal dashed line shows the total number of HPKs in each genome. Symbols indicate the evolutionary distance (arbitrary units, see Methods) of outgroups used in the analysis. Some clades with greater taxon sampling have better resolved timings. The last panel shows average numbers for all genomes.

DOI: [10.1371/journal.pcbi.0020143.g011](https://doi.org/10.1371/journal.pcbi.0020143.g011)